7.1 The Stanford Classifier
7.2 Others
8. A lightweight, accurate classifier
9. Summary of conclusions

1 Overview

This section introduces two classifier models, Naive Bayes and Maximum Entropy, and evaluates them in the context of a variety of sentiment analysis problems. Throughout, I emphasize methods for evaluating classifier models fairly and meaningfully, so that you can get an accurate read on what your systems and others' systems are really capturing.

Demo Trained classifier models to experiment with:
http://sentiment.christopherpotts.net/classify/

2 Models

I concentrate on two closely related probabilistic models: Naive Bayes and MaxEnt. Some other classifier models are reviewed briefly below as well.

2.1 Naive Bayes

The Naive Bayes classifier is perhaps the simplest trained, probabilistic classifier model. It is remarkably effective in many situations.

I start by giving a recipe for training a Naive Bayes classifier using just the words as features:

1. Estimate the probability $P(c)$ of each class $c \in C$ by dividing the number of words in documents in c by the total number of words in the corpus.
2. Estimate the probability distribution $P(w \mid c)$ for all words w and classes c. This can be done by dividing the number of tokens of w in documents in c by the total number of words in c.

3. To score a document *d* for class *c*, calculate

$$\textbf{score}(d,c) \stackrel{def}{=} P(c) * \prod_{i=1}^{n} P(w_i|c)$$

4. If you simply want to predict the most likely class label, then you can just pick the c with the highest **score** value. To get a probability distribution, calculate

$$P(c|d) \stackrel{def}{=} \frac{\textbf{score}(d,c)}{\sum_{c' \in C} \textbf{score}(d,c')}$$

The last step is important but often overlooked. The model predicts a full distribution over classes. Where the task is to predict a single label, one chooses the label with the highest probability. It should be recognized, though, that this means losing a lot of structure. For example, where the max label only narrowly beats the runner-up, we might want to know that.

The chief drawback to the Naive Bayes model is that it assumes each feature to be independent of all other features. This is the "naive" assumption seen in the multiplication of $P(w_i \mid c)$ in the definition of **score**. Thus, for example, if you had a feature `best` and another `world's best`, then their probabilities would be multiplied as though independent, even though the two are overlapping. The same issues arise for words that are highly correlated with other words (idioms, common titles, etc.).

2.2 Maximum Entropy

The Maximum Entropy (MaxEnt) classifier is closely related to a Naive Bayes classifier, except that, rather than allowing each feature to have its say independently, the model uses search-based optimization to find weights for the features that maximize the likelihood of the training data.